Linguistics Data Interest Group Meeting @RDA Plenary 10

Tuesday September 19, 2017 4-5:30 (Mansfield 9)

This document is no longer being edited
Contact the LDIG group (Co-chair Lauren Gawne l.gawne@latrobe.edu.au)

Session page: https://www.rd-alliance.org/ig-linguistics-data-rda-10th-plenary-meeting

Agenda

4-4:05: An introduction to the LDIG, goals, and directions
4:05-4:15: A high-level introduction to Linguistics data issues and starting points
4:15-4:25: Overview of the Austin principles and the motivation behind them with summary of current feedback
4:25-4:40: Feedback from floor, comments from DCWG and others
4:40-5:10: Real-time text updates and drafting
5:10-5:30: How to proceed next, forming a WG if necessary

Relevant documents for this meeting

- Austin Principles of Data Citation in Linguistics
- Feedback received on Austin Principles (feel free to add yours!)
- LDIG RDA page
- LDIG charter statement
- View the slides for this meeting.

Introduction to LDIG, Linguistic Data & the Austin Principles

See the slides for information

Time for Feedback from floor

Information and comments added directly to the text below.

Real-time updating of Austin Principles text

DRAFT June 6 2017

About this document

The Austin Principles of Data Citation in Linguistics is currently being developed by the participants in the *Data Citation and Attribution for Reproducible Research in Linguistics* project, in cooperation with the Research Data Alliance Linguistics Data Interest Group. The content of this document is based on the FORCE11 Joint Declaration of Data Citation Principles, specifically annotated for the discipline of linguistics. Original text from the FORCE11 document is in grey plain font, and linguistics-related annotations are in *blue italics*.

These principles are intended to provide guidance to the field in developing specific formats for formatting citations. We strongly encourage journal editors and publishing companies to develop a comprehensive stylesheet of formats for authors to use when citing digital data sets. However, in the absence of such stylesheets, we urge individual authors and researchers to adapt the principles below for their own citations to the best of their ability.

Your comments on this draft document are welcome; please email us to contribute your thoughts.

Please cite this document as:

Data Citation and Attribution in Linguistics Group. 2017. *Austin Principles of Data Citation in Linguistics, Draft June 6, 2017*. https://sites.google.com/a/hawaii.edu/data-citation/austin-principles-of-data-citation-in-linguistics.

PREAMBLE

Data, in all its many varieties of shapes and formats, are fundamental to science, including the field of linguistics, and should be treated as such. From the Joint Declaration of Data Citation Principles,

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

In support of this assertion, and to encourage good practice, we offer a set of guiding principles on data citation for linguists who make reference to data within scholarly literature, another dataset, or any other research object. These principles are based on the FORCE11 Joint Declaration of Data Citation Principles, annotated specifically for the field of linguistics and all of its subfields.

PRINCIPLES

The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.
These citation principles are not comprehensive recommendations for data stewardship. And, as practices vary across communities and technologies will evolve over time, we do not include recommendations for specific implementations, but encourage communities to develop practices and tools that embody these principles.
The principles are grouped so as to facilitate understanding, rather than according to any perceived criteria of importance.

1. Importance

Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

In linguistics, data form not only a record of scholarship, but of cultural heritage, societal evolution, and human potential. Because of their importance in these areas, linguistic data are of fundamental importance to the field and should be treated as such.

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

In linguistics, this applies not only to the researchers, but (when appropriate and possible) any individuals who participate in the collection or creation of those data, including native speakers, interviewees, and transcribers.

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
In linguistics, the method of data collection should also be made apparent in the text, e.g. a native speaker judgment, recorded audio, written excerpt, ethnographic notes.

4. Unique Identification

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

In linguistics, many data repositories specializing in linguistic data, like DELAMAN archives and TROLLing, offer such identification in the form of a Persistent Identifier (PID), such as Digital Object Identifier (DOI), or Handle.

[Insert an example here]

5. Access

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
Data should be as open as possible and as closed as necessary based on relevant ethical, legal and speaker community constraints. Researchers should strive wherever possible to make their data open in research protocols rather than closed.

6. Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

Linguists should confirm that the archives or repositories where they are storing their data have written policies pertaining to persistence of data and metadata.

7. Specificity and Verifiability

Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
For data uses that require a fine-grained citation for clarity, a systematic method of identification for the data should be used.
Sherzer, Joel (Researcher), Lanni (Contributor), Olowiktinappi (Performer), Armando Gutiérrez (Translator). (1970). "Myth of White Prophet - complete version" CUK001R002I200.pdf page 2. Kuna Collection of Joel Sherzer. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: text. Access: public. Resource: CUK001R002.

8. Interoperability and Flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Next steps/other thoughts

- Booth at LSA and other conferences.
- Anecdotes about data that has been reused leading to new discoveries. Send out a request for data reuse success stories.
- Think about the disciplinary cross-over wrt the content of narratives etc.
- How to reach out to the rest of our audience (the ten-legged race) - journals, researchers, societies, etc.

- We can explain ourselves to Grand Challenges of RDA, why we are relevant.
- “Connecting with the macro-level discourses” - can we catch the attention of funders, to motivate people and other levels.
- IPY is an example: start with why language data is important to the whole world, then explain why we need better data work - 1000 people came to the International Data Week last week. If our community doesn’t get behind this, we’ll be left behind in 5 years. What’s our “go to the moon”?
- Everyone is more aware of language now (siri, social media).
- Lynn: carrots and sticks very by community. In earth sci, it’s cv items and extra money for funding. We should be proactive to set our own standards, those become drivers from the bottom up.
- Joel: software needs to make citation easier, then buy-in will be easier.

Relevant documents for this meeting

- Austin Principles of Data Citation in Linguistics
- Feedback received on Austin Principles (feel free to add yours!)
- LDIG RDA page
- LDIG charter statement
- View the slides for this meeting.